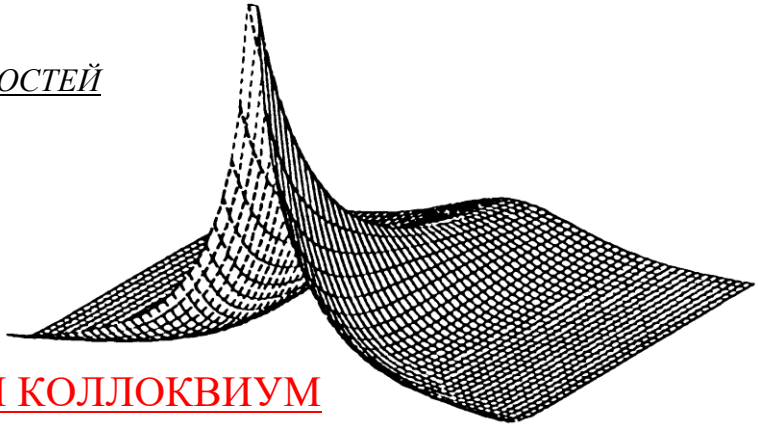
Кафедра ТЕОРИИ ВЕРОЯТНОСТЕЙ



АСПИРАНТСКИЙ КОЛЛОКВИУМ

С марта 2017 года на кафедре теории вероятностей начинает работать *аспирантский коллоквиум* по теории вероятностей, математической статистике, теории случайных процессов. На заседаниях будут заслушиваться доклады аспирантов (по текущей работе) и преподавателей (по специальным и общим вопросам).

Аттестация аспирантов будет коррелироваться с их участием в этом коллоквиуме.

15 марта 2017 г.

Аспирант профессора А.В. Булинского
А.С. Ракитько

Последовательный отбор переменных в MDR-EFE методе

Задача выявления факторов, объясняющих некоторый случайный отклик Y, возникает во многих прикладных исследованиях. Например, в медико-биологических исследованиях в качестве факторов могут выступать генетические маркеры (SNP): $X1,\ldots,Xn$, принимающие дискретные значения, а Y показывает наличие или отсутствие заболевания (соответственно, значения 1 или −1). Как правило, число исследуемых факторов n много больше количества имеющихся N наблюдений. При этом считается, что количество значимых факторов $Xk1,\ldots,Xkr$, влияющих на Y , невелико. Для поиска таких факторов применяется различная техника (LARS, LASSO, MDR, Bayes analysis и другие).

Нами используется метод MDR-EFE (*Multifactorial Dimensionality Reduction with Error Function Estimation*) понижения размерности набора факторов. Этот метод основан на анализе статистической оценки функционала ошибки предсказания отклика. Данный функционал задается формулой $Err(fP\ A) = |Y - f(X)|\psi(Y\ )$, где f – предсказательный алгоритм, а $\psi(\cdot)$ – штрафная функция. В предыдущих работах был установлен критерий сильной состоятельности предлагаемых оценок функционала ошибки, а также доказаны варианты центральной предельной теоремы для введенных статистик. Метод применим к широкому классу моделей, но имеет высокую вычислительную сложность, поскольку приходится перебирать большое число комбинаций факторов. Для упрощения и ускорения алгоритма предлагается использовать последовательный отбор переменных (*forward selection*). Предположение модели наивного байесовского классификатора приводит к некоторой модели логистической регрессии и позволяет дать нижнюю оценку вероятности события A, состоящего в правильном нахождении r значимых факторов при описанном выше алгоритме последовательного поиска. Кроме того, рассмотрен вариант алгоритма с регуляризованной версией функционала ошибки.

Заседания будут проводиться по средам в аудитории 16-03 Главного Здания Москвоского Государственного Университета имени М.В. Ломоносова с 18:10 до 19:10